A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning

Atsuo Suemitsu, Jianwu Dang, Takayuki Ito, Mark Tiede

A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning

Atsuo Suemitsu[a] **and Jianwu Dang**
School of Information Sciencen, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
sue@jaist.ac.jp, jdang@jaist.ac.jp

Takayuki Ito
CNRS, GIPSA-Lab, 11 Rue des Mathématiques, 38402 Saint Martin, d'Heres Cedex, France
takayuki.ito@gipsa-lab.grenoble-inp.fr

Mark Tiede
Haskins Laboratories, New Haven, Connecticut 06511, USA
tiede@haskins.yale.edu

Abstract: Articulatory information can support learning or remediating pronunciation of a second language (L2). This paper describes an electromagnetic articulometer-based visual-feedback approach using an articulatory target presented in real-time to facilitate L2 pronunciation learning. This approach trains learners to adjust articulatory positions to match targets for a L2 vowel estimated from productions of vowels that overlap in both L1 and L2. Training of Japanese learners for the American English vowel /æ/ that included visual training improved its pronunciation regardless of whether audio training was also included. Articulatory visual feedback is shown to be an effective method for facilitating L2 pronunciation learning.

[DOS]
Date Received: May 26, 2015 **Date Accepted:** September 7, 2015

1. Introduction

In second language (L2) acquisition, learners continue to have difficulty in achieving native-like production even if they receive instruction on how to correctly position the speech articulators. The essential issue is that the learners lack appropriate knowledge on how to modify their articulation to produce correct L2 sounds.[1,2] In this situation, acquiring appropriate patterns of articulation for correct L2 production may also be affected by their native language (L1), that is, L1 interference.[3,4]

Training with real-time visual feedback of articulatory position using ultrasound imaging[2,5] or electromagnetic articulometry (EMA)[6,7] has been proposed to address this issue. These approaches have improved the production of English approximants[2] and of French vowels[5] by native Japanese speakers and Japanese flap[6] by native American English (AE) speakers. However, because the previous studies had employed audio stimuli and/or provided knowledge of phonation during training, it is unclear whether or not the visual presentation of articulatory information itself actually plays a main role in facilitating L2 pronunciation learning. In addition, these articulatory training systems require a teacher or clinician to provide a model for learners to follow; this may be problematic in some circumstances.

The visualization of target articulatory position or posture can assist because the learners can make direct use of feedback in adjusting their articulators for the correct pronunciation. However, given that vocal tract shape differs in individuals, the estimation of target articulatory positions appropriate for each learner is a challenging task.

EMA has the potential for presenting an articulatory target together with real-time visual feedback of current articulator position because quantitative data for the position and velocity of sensors characterizing articulatory motion are available in real time. However, apart from Levitt and Katz,[6] who tracked a single target point for

[a]Author to whom correspondence should be addressed.

the acquisition of Japanese flaps, no previous study has presented multiple targets (points or shape) for learning novel sounds in an L2 context.

This study proposes an EMA-based real-time articulatory visual feedback approach that can provide target articulatory positions for application to vowels as well as consonants. We examine whether our proposed approach improves pronunciation by Japanese learners of AE by conducting within-session training for pronunciation of the AE vowel /æ/, which is not part of the typical Japanese five vowel inventory. To investigate the effect of real-time articulatory visual feedback, we compare three conditions: presentation of acoustic information only, articulatory information only, and presentation of both. It is expected that articulatory-based training with target presentation can assist in overcoming a variety of difficulties, such as L1 interference, in L2 pronunciation learning.

2. EMA-based, real-time, visual feedback system

We designed a system using an EMA as shown in Fig. 1(a) to present articulatory positions in real time together with the target articulatory positions estimated from speaker acoustics and articulatory data. The three-dimensional (3D) EMA system (AG500, Carstens Medizinelektronik) tracks positions of sensors glued to the speech articulators and reference points. Sensors were placed as shown in Fig. 1(b): the tongue tip (TT), blade (TB), dorsum (TD), lower incisors (LI), upper lip (UL), and lower lip (LL), together with reference sensors on the upper incisors, nasion, and mastoid processes tracked to compensate for head movement (all midsagitally placed apart from the mastoid references). Articulatory movement and speech acoustics are digitized at sampling rates of 200 Hz and 16 kHz, respectively. For visualization, sensor position data were transformed to a coordinate system based on each participant's occlusal plane and corrected for head movement. Acquisition and transformation of the data were repeated every 50 ms, that is, the articulatory presentation was updated at a 20 Hz rate. No perceptually apparent latency between sensor motion and its visualization was observed. Figure 1(c) shows an example of the real-time visual feedback display. The tongue surface contour was obtained using cubic spline interpolation through the three tongue sensor positions.

The speaker-specific target articulatory positions for /æ/ were estimated using a multiple linear regression model trained on native AE speakers, using as predictors each learner's acoustic and kinematic productions of AE vowels /a/, /i/, and /u/, which overlap reasonably well with Japanese /a/, /i/, and /ɯ/. This approach enabled us to provide an estimate of /æ/ target position compatible with each learner's particular vocal tract shape.

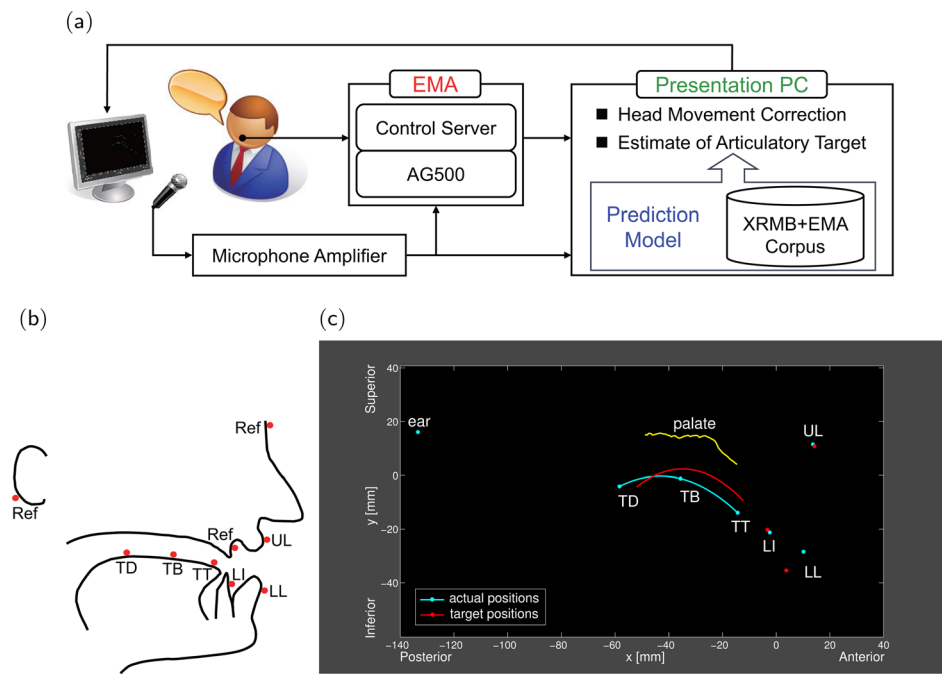

Fig. 1. (Color online) (a) Schematic of the developed system. (b) Sensor placement. (c) Example of the real-time visual feedback display.

The prediction model was constructed using the acoustic and kinematic data of 49 native AE speakers from the University of Wisconsin x-ray microbeam (XRMB) speech production corpus[8] (19 males and 21 females), augmented by EMA data (5 males and 4 females) collected at Haskins Laboratories. Specifically, 12 models [6 articulatory attachment points (TT, TB, TD, LI, UL, and LL) × 2 coordinates (*x,y*)] were built using stepwise selection of these predictors: the first (F1) and second formant (F2) frequencies; posterior/anterior (X) and inferior/superior (Y) coordinate values of TT, TB, TD, LI, UL, and LL for AE vowels /a/, /i/, and /u/; the area of the triangle defined by these vowels in the F1 × F2 plane; and the area of the triangle defined by the TDxy positions associated with these vowels. Here TB was estimated by calculating a midpoint between T2 (mid-ventral) and T3 (mid-dorsal) pellets from the Wisconsin XRMB corpus, and TT and TD corresponded to T1 (ventral) and T4 (dorsal) pellets, respectively.

3. Methods

The participants were 21 male monolingual native speakers of Japanese, aged between 22 and 35 yr, with no self-reported hearing deficits or speech disorders. All participants had received some English instruction in school but had no overseas living experience. The Ethical Committee of the Japan Advanced Institute of Science and Technology (JAIST) approved the experimental procedures, and all participants provided written informed consent.

The experimental sequence consisted of four phases: preparation, pre-test, training, and post-test. In the preparation phase, sensors were attached to the speech articulators as shown in Fig. 1(b), palatal shape was measured for real-time visualization, and the occlusal plane was sampled to provide a consistent coordinate system during real-time display. In addition, the articulatory positions and speech data for the participant's production of the sustained Japanese vowels /a/, /i/, and /ɯ/ were collected for estimation of that speaker's articulatory /æ/ position. The additional Japanese vowels /e/ and /o/ were also recorded but not used for modeling.

Participants were tested under one of three contrasting experimental conditions: visual feedback of tongue position with no acoustic cue (V condition), acoustic cue with no visual feedback (A condition), and visual feedback presented with acoustic cue (VA condition). We tested six subjects in V condition, seven subjects in A condition, and eight subjects in VA condition, respectively. For the V condition, a target word was presented on a computer screen with no audio cue. For the A and VA conditions the on-screen target word was co-presented with an audio "cue" production of the target by a male AE speaker selected from the XRMB corpus.

The target words were "back," "sad," and "had," or the vowel /æ/ produced in isolation. In the pre- and post-test phases of the experiment, each word was presented five times in randomized order, elicited by stimuli tailored to the condition under test. Each participant, fitted with 10 EMA sensors attached as shown in Fig. 1(b), was seated in front of a computer screen in a quiet room at JAIST. Articulatory recordings were collected using an EMA (AG500) at 200 Hz synchronized with concurrent audio recorded with a directional microphone (NTG-3, RODE) at 16 kHz.

In the training phase of the V and VA conditions, the participant was first asked to fit his/her tongue contour and the displayed UL, LL, and LI positions to the estimated target positions without producing speech sounds for about 5 min to facilitate learning motor control of this novel articulation without phonation. Then the participant practiced the production of the vowel /æ/ after matching his/her articulation to the target as elicited by visual and/or audio stimulus presentation. This task was repeated 20 times. In the A condition, the participant was asked to try to imitate the vowel /æ/ as elicited by audio cue. This task was also repeated 20 times.

To assess the effect of individual training, the F1 and F2 of the vowel /æ/ were obtained from the acoustic recordings from the pre- and post-test phases. Formant estimates were obtained using a 14th order LPC analysis using a 0.9 pre-emphasis factor and a Hanning window length of 25 ms with 15 ms overlap between frames. Using the corresponding spectrogram to verify spectral stability, five frames taken from the approximate center of each utterance were analyzed and the resulting formants were averaged. Additionally, taking the human auditory system and the anisotropy of the F1 × F2 frequency plane into account, all formant values were converted to the equivalent rectangular bandwidth (ERB) units.[9] The difference between conditions was evaluated by calculating the Euclidean distances between the produced sounds and a reference sound, established as the median of the native AE speaker productions in the F1 × F2 ERB space. A linear mixed model (LMM) analysis was

applied with condition (A, V, VA) and phase (pre, post) as fixed factors including their interaction and by-participant intercepts as random effects, using the lme4 and lmerTest packages within R (www.R-project.org).

For the articulatory analysis, we characterized the articulatory positions of the utterance /æ/ by averaging the data points over the identical measurement period used in the acoustic analysis. To quantify the change in articulatory position between pre- and post-test phases, the Euclidean distance between the pre- and post-test positions of each sensor was calculated after averaging the coordinate values for each sensor over the measured intervals. The difference (post- minus pre-test) of each position between the pre- and post-test phases was then computed to investigate how each articulator was affected by the training.

4. Results and discussion

Figure 2(a) shows the distribution of the produced /æ/ sounds in the F1 × F2 space for each condition at pre- (crossmark) and post-test (circle) phases for representative participants. Central ellipses represent the 95% confidence limits for the /æ/ distribution for the male AE speakers obtained from the XRMB and EMA corpus, and vertical and horizontal lines indicate the medians of the F1 and F2 values of the native /æ/ distribution, respectively, and dashed and dotted ellipses are the 95% confidence limits for the Japanese /e/ and /a/ distributions for all participants, respectively. We found that the produced /æ/ sounds after training were distributed closer to the center of the native /æ/ distribution in the V and VA conditions but not in the A condition. Figure 2(b) compares the average acoustic distance from target at the pre- and post-test for each participant under three conditions where error bars represent standard error of the mean ($n = 20$). The mean values for each condition are shown in Fig. 2(c), and the results of the LMM analysis (840 observations from 21 participants) are reported in Table 1. The estimates of fixed effects showed greater acoustic distance from the target for the V condition from the A baseline ($t = 2.44$, $p < 0.03$), consistent with the absence of audio stimulus in that condition. There was no effect of phase overall, but both conditions with visual training (V, AV) showed highly significant reductions in acoustic distance at the post-test phase following training ($t = -5.23$, $p < 0.0001$; $t = -4.06$, $p < 0.0001$). These results suggest that the real-time visual feedback of articulatory information with a target facilitates improvements in pronunciation of the non-native vowel for Japanese learners.

Figure 3(a) shows the articulatory distribution of pre- and post-test utterances for the same participants as in Fig. 2(a), where each ellipse represents the 95% confidence limits for the pre- or post-test utterances at each articulator. We found that the articulatory positions between pre- and post-test phases in the V and VA conditions were affected more than those in the A condition. Figures 3(b) and 3(c) show the

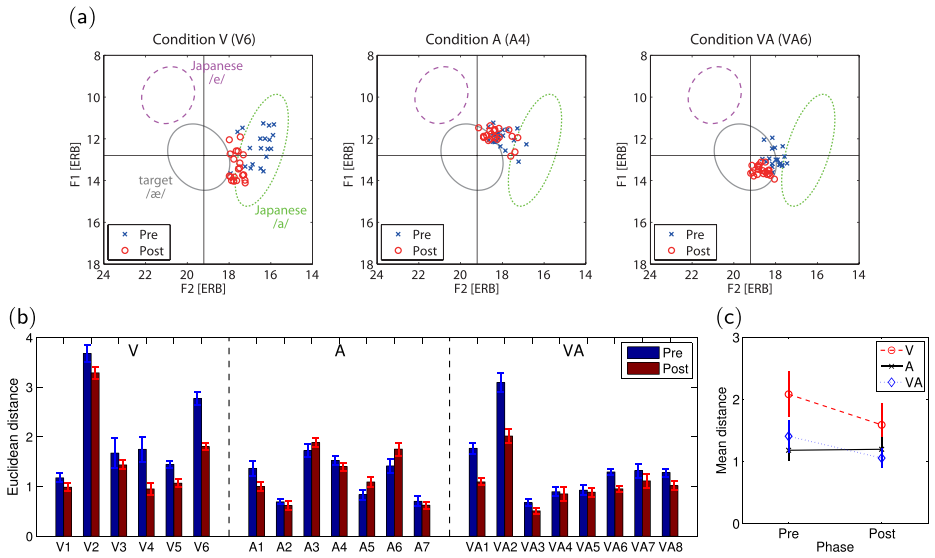

Fig. 2. (Color online) (a) Scatterplots of utterances of representative participants in the F1 × F2 space at pre- and post-test phases for (left) V, (middle) A, and (right) VA conditions. (b) Euclidean distance between the produced and reference sounds at the pre- and post-test phases for each participant. (c) Average distance over all participants for each condition.

Table 1. Results of the LMM predicting acoustic distance from conditions and phases. SD, standard deviation; SE, standard error; df, degree of freedom.

	Random effects			Fixed effects				
	Variance	SD		Estimate	SE	df	*t*-value	*p* value
Participant	0.43	0.65	(Intercept)	1.18	0.25	18.6	4.70	0.00
Residual	0.31	0.55	Condition V	0.90	0.37	18.6	2.44	0.02
			Condition AV	0.23	0.34	18.6	0.66	0.52
			Phase post	0.02	0.07	816	0.24	0.81
			Condition V:Phase post	−0.51	0.10	816	−5.23	0.00
			Condition VA:Phase post	−0.37	0.09	816	−4.06	0.00

average Euclidean distance and difference across all participants for each condition, respectively. In general, the shift amount of the tongue, jaw, and lower lip in the V and VA conditions was greater than that in A condition [Fig. 3(b)]. In particular, the tongue was displaced anteriorly and inferiorly, and jaw and mouth were more open in both V and VA conditions [Fig. 3(c)]. Because the articulatory positions of /æ/ were originally located between those for the Japanese vowels /a/ and /e/ at the pre-test phase, the larger position changes post-test indicate that the learners' articulation moved closer to the correct articulatory position for the production of the AE vowel /æ/. In contrast, the small displacement in the A condition was not shifted toward the articulation of /æ/ because the tongue was displaced differently in comparison with the V and VA conditions [Fig. 3(c)]. These results suggest that the real-time visual feedback plus target presentation can induce motor learning appropriate for improvement in the production and pronunciation of /æ/.

In summary, the data suggest a short-term learning effect induced in the V and VA conditions but not in the A condition. This might be due to L1 interference.[3,4] When adult learners learn an unfamiliar L2 pronunciation from acoustic cues alone (as in A condition), they will utilize their existing knowledge of L1 or the learning experience of L2 to reproduce the L2 sound being aimed at. Thus on the one hand, they may prefer to use a motor command based on their existing acoustic-articulatory mapping rather than to generate a new motor command specific to the new L2 sound. On the other hand, when adult learners are asked to learn a new motor command by presenting the articulatory target for L2 pronunciation, as in the V and VA conditions, they can exploit the additional feedback information directly, minimizing the effects of L1 interference on the acquisition of the new motor command. Consequently, they will be able to produce a more accurate L2 sound using the newly acquired motor command.

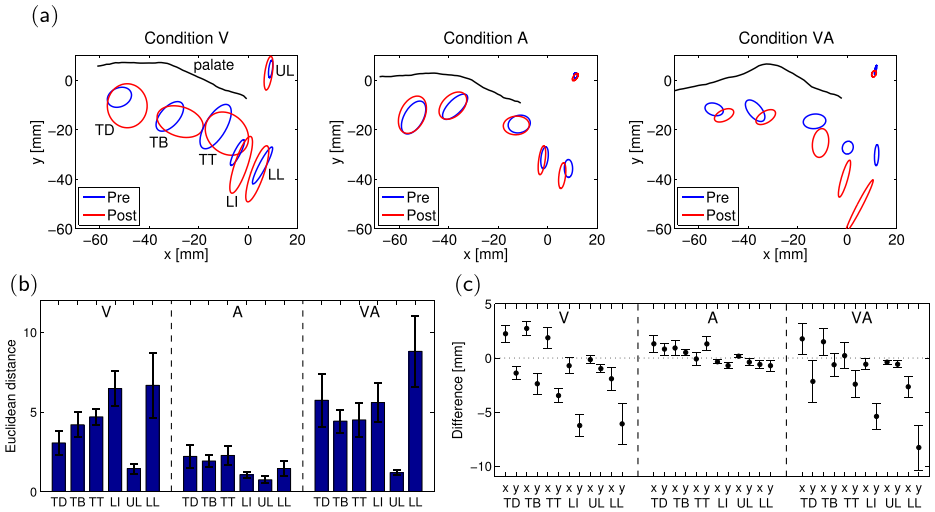

Fig. 3. (Color online) (a) Articulatory distribution at the pre- and post-test phases for the participants in Fig. 2(a). (b) Euclidean distance between the pre- and post-test articulatory positions for each condition. (c) Difference (post- minus pre-test mean coordinates of each articulatory position) per condition.

5. Concluding remarks

We have proposed an EMA-based real-time articulatory visual feedback approach that provides a speaker-specific target presentation. Short-term training with the proposed approach improved the pronunciation of Japanese learners in acquiring the non-native vowel /æ/. We demonstrated that independent of audio presentation, real-time visual feedback plus target presentation can improve articulatory positions associated with correct pronunciation, leading to associated improvements in produced audio. This suggests that short-term articulatory training can be a useful tool for overcoming a variety of difficulties, such as L1 interference, in L2 pronunciation learning. The findings of the present study are preliminary; further studies are required to examine the retention effect of the improved pronunciation and to evaluate the quality of the pronunciation through perceptual evaluation by native AE speakers.

Acknowledgments

This work was partly supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Nos. 25240026, 25330190, and 25370444). Additional support was provided by National Institutes of Health Grant No. DC-002717.

References and links

[1]P. Badin, A. Ben, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *Interspeech 2010 Satellite Workshop*, Tokyo (2010), Actes de SLATE, pp. P1–10.

[2]B. Gick, B. M. Bernhardt, P. Bacsfalvi, and I. Wilson, "Ultrasound imaging applications in second language acquisition," in *Phonology and Second Language Acquisition*, edited by J. G. Hansen Edwards and M. L. Zampin (John Benjamins, Amsterdam, 2008), pp. 309–322.

[3]S. Gass and L. Selinker, *Second Language Acquisition: An Introductory Course*, 3rd ed. (Erlbaum, Hillsdale, NJ, 2008).

[4]C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*, edited by M. J. Munro and O. S. Bohn (Benjamins, Amsterdam, 2007), pp. 13–34.

[5]C. Pilot-Loiseau, T. K. Antolik, and T. Kamiyama, "Contribution of ultrasound visualisation to improving the production of the French /y/-/u/ contrast by four Japanese learners," in *Proceedings of Phonetics, Phonology and Languages in Contact*, Paris (2013), pp. 86–89.

[6]J. S. Levitt and W. F. Katz, "The effect of EMA-based augmented visual feedback on the English speakers' acquisition of the Japanese flap," in *ISCA Proceedings of Interspeech*, Makuhari (2010), pp. 1862–1865.

[7]W. Katz, T. Campbell, J. Wang, E. Farrar, J. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-Speech: A real-time, 3D visual feedback system for speech training," in *ISCA Proceedings of Interspeech*, Singapore (2014), pp. 1174–1178.

[8]J. R. Westbury, *X-Ray Microbeam Speech Production Database User's Handbook* (University of Wisconsin, Madison, WI, 1994).

[9]B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1–2), 103–138 (1990).